Applications of single-cell and bulk RNA sequencing in onco-immunology

Maria Kuksin, Daphné Morel, Marine Aglave, François-Xavier Danlos, Aurélien Marabelle, Andrei Zinovyev, Daniel Gautheret, Loïc Verlingue

Version of Record: https://www.sciencedirect.com/science/article/pii/S0959804921001532
Manuscript_cc0a6204602600ba0a024f68f9fefb27

Applications of single-cell and bulk RNA sequencing in onco-immunology

Maria Kuksin[1,2+], Daphné Morel[2,3,4+], Marine Aglave[5], François-Xavier Danlos[6], Aurélien Marabelle[2,6], Andrei Zinovyev[7,8,9,10], Daniel Gautheret[11,12,13], Loïc Verlingue[2,4,7,13 *]

[1]ENS de Lyon, 15 Parvis René Descartes, 69007 Lyon, France ; [2]Département d'Innovations Thérapeutiques et Essais Précoces (DITEP), Gustave Roussy Cancer Campus, 114 rue Edouard Vaillant, 94800 Villejuif, France ; [3]Département de radiothérapie, Gustave Roussy Cancer Campus, Gustave Roussy, 114 rue Edouard Vaillant, 94800 Villejuif, France ; [4]INSERM UMR1030, Molecular Radiotherapy and Therapeutic Innovations, Gustave Roussy, 114 rue Edouard Vaillant, 94800 Villejuif, France ; [5]INSERM US23, CNRS UMS 3655, Gustave Roussy Cancer Campus, 114 rue Edouard Vaillant, 94800 Villejuif, France ; [6]INSERM U1015, Gustave Roussy, Université Paris Saclay ; [7]Institut Curie, PSL Research University, F-75005 Paris, France; [8]INSERM, U900, F-75005 Paris, France ; [9]MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France ; [10]Laboratory of advanced methods for high-dimensional data analysis, Lobachevsky University, 603000 Nizhny Novgorod, Russia ; [11]Institute for Integrative Biology of the Cell, UMR 9198, CEA, CNRS, Université Paris-Saclay, Gif-Sur-Yvette, France ; [12]IHU PRISM, Gustave Roussy Cancer Campus, Gustave Roussy, 114 rue Edouard Vaillant, 94800 Villejuif, France ; [13]Université Paris-Saclay, France.

[+] These authors contributed equally to the work

***Corresponding author:** loic.verlingue@gustaveroussy.fr – tel. +33142115591 – 114 rue Edouard Vaillant, 94800 Villejuif, France

Word counts

- 157-words abstract
- 5,189-words core text
- 4 figures, 1 table, 4 boxes
- 148 references

Highlights

- Personalization of onco-immunotherapy treatment relies on both cancer and immune profiling.
- Single-cell RNA-seq delivers detailed but sparse information.
- Bulk RNA-seq requires extrapolation to quantitatively estimate the immune contexture.
- We reviewed standard methods that can boost clinical applications of tumor RNA-seq.

Abstract

The rising interest for precise characterization of the tumor immune contexture has recently brought forward the high potential of RNA sequencing (RNA-seq) in identifying molecular mechanisms engaged in the response to immunotherapy. In this review, we provide an overview of the major principles of single-cell and conventional (bulk) RNA-seq applied to onco-immunology. We describe standard pre-processing and statistical analyses of data obtained from such techniques and highlight some computational challenges relative to the sequencing of individual cells. We notably provide examples of gene expression analyses such as differential expression analysis, dimensionality reduction, clustering and enrichment analysis. Additionally, we used public datasets to exemplify how deconvolution algorithms can identify and quantify multiple immune subpopulations from either bulk or single-cell RNA-seq. We give examples of machine and deep learning models used to predict patient outcomes and treatment effect from high-dimensional data. Finally, we balance the strengths and weaknesses of single-cell and bulk RNA-seq regarding their applications in the clinic.

Key words

Immunotherapy; single-cell RNA-seq; bulk RNA-seq; tumor microenvironment; cancer; precision medicine

Introduction

Tumor samples always contain heterogeneous cell populations that comprise – beyond malignant cells – stromal and immune tumor-associated cells (Figure 1) [1]. After a long period of focusing on the tumor's genetic alterations, we have slightly entered an era in which the behavior of non-tumoral cells becomes as meaningful as the tumor itself. Immunotherapy guidance for first line treatment and beyond still lacks reliable biomarkers, yet accumulating evidences point towards the importance of studying a tumor tissue as a whole to better capture its potential of response to immune-checkpoint blockers: exhaustive composition, density, functional state, organization and interaction between cellular subtypes, including tumor cells [2,3].

Tumor immune infiltrates can associate CD8+ cytotoxic T cells, CD4+ helper cells (T_H1 cells), dendritic cells, CD25+/FOXP3+ regulatory T cells, myeloid-derived suppressor cells (MDSCs), natural-killer cells, B lymphocytes, tumor-associated macrophages, among others. Some of those promote immune escape of the tumor (e.g. MDSCs), some elicit anti-tumoral actions (e.g. dendritic cells), others can either promote or prevent immune escape depending on their activation status (e.g. CD8+ T cells) [4,5]. For example, the presence and proportion of M1 macrophages, B cells and T_H1 lymphocytes within the immune infiltrate almost always correlates with good prognosis, whereas the presence of regulatory T cells or M2 macrophages usually associates with bad prognosis [3]. Even though their cellular behavior may be influenced by the tumor, it is expected that those non-tumoral immune cells share the same DNA sequence, which is specific to the host (with the exceptions of rearranged immunoglobulin and TCR loci). However, they significantly vary phenotypically and harbor separate mRNA expression profiles, which makes them identifiable by gene expression analysis.

Among various techniques deployed to assess gene expression [6], RNA sequencing (RNA-seq) can provide qualitative (RNA sequence) and quantitative (RNA abundance) analyses of

either targeted mRNA transcripts or the complete transcriptome of a particular tissue [7–9]. Two methods of RNA-seq are henceforth commonly considered for onco-immunology studies: standard bulk RNA-seq and single-cell RNA-seq (scRNA-seq). Standard bulk RNA-seq started replacing microarray techniques in the late 2000s and is now widely used in translational research. For each transcript, it provides an average expression level in the sample, which may comprise different cell types [10]. Conversely, scRNA-seq is a relatively new technology that measures the gene expression levels for each transcript within each individual cell of the sample and allows a representation of the distribution of this expression in each sub-population of cells [11,12]. Therefore, scRNA-seq rather aims to study a particular cell or cell type behavior in the specific context of its microenvironment, while bulk RNA-seq was not primarily designed for the precise characterization of a tumor composition [13].

At first glance, one could believe that scRNA-seq represents the most suited method to draw a complete picture of a tumor and thus, may be a method of choice to study the immune tumor-related features that could be predictive of response to immunotherapy. Yet, the initial substantial excitement regarding scRNA-seq theoretical promises has historically faced technical limitations [13]. In addition, it is now technically possible to extrapolate the cellular composition of a tumor sample from bulk analysis, which at least partially alleviates the conceptual frontier between bulk and scRNA-seq full potentials.

Here, we give an overview of some standard methods used for processing single-cell or bulk RNA-seq data. We describe how scRNA-seq and bulk methods can share several computational analysis concepts (reads quality control, mapping or quasi-mapping, differential expression analysis, dimensionality reduction, clustering, etc.) – sometimes with the same objectives, sometimes with very distinct outputs – bearing in mind that for every task, scRNA-seq analysis will often imply additional dimensionalities to estimate one complete gene expression profile per individual cell.

We notably emphasize the statistical approaches that can be used to study the tumor microenvironment composition – including from bulk analysis – in a qualitative and quantitative manner. We also describe machine learning tools that aim at predicting survival or tumor response following treatment with immune-checkpoint blockers. Finally, we depict the recent evolution in the clinical development field of onco-immunology that increasingly integrates RNA-seq in clinical trial designs, reflecting the growing interest in this approach for multiple purposes.

Single-cell and bulk RNA-seq processing: basic clues to appreciate the differences

Collecting, storing and processing RNA

Unlike DNA, which is a stable structure, RNA is highly sensitive to oxidation and hydrolytic cleavage of its phosphodiester bonds by nucleases, and therefore requires conscientious storage to prevent rapid degradation [14]. Bulk RNA-seq can be performed on human tumor samples obtained from surgical resection or biopsies, preserved in formalin-fixed and paraffin embedded (FFPE) or freshly frozen [15–17]. Thus, all the samples can be collected in due course and processed all in the same time.

Additional facts to consider for scRNA-seq

On the other hand, scRNA-seq analysis requires viable cells suspension to allow robust cell isolation, which contraindicates any fixation of freezing steps. ScRNA-seq therefore requires processing the samples quite promptly – within a few days – after the time of collection. Another important point to consider is that during that process, the cellular viability of each sub-population of cells can be differently affected, leading to some cell types being over- or under-represented in the final dataset (source of interpretation bias). Therefore, before planning a scRNA-seq analysis, several questions need to be addressed, including: (i) which and how many cell types are to be analyzed (for example, only tumor cells, all stromal

immune cells or only a subset of immune cells such as memory CD4+ T cells or sparse population of a subset of dendritic cells), (ii) which method is used to prepare cell suspension (tissue dissociation, selected cells enrichment and cells conservation), (iii) which single-cell technology (droplet- or plate-based; short or long reads; Unique Molecular Identifiers or not) and (iv) which library and other RNA-seq parameters to use [18–21].

Sequencing and basic bias correction

The following steps of single-cell and bulk RNA-seq usually comprise reverse transcription, reads sequencing, mapping or quasi-mapping, data format transformation, count table generation and inference, such as the exploration of gene expression and/or transcript fusions (Figure 2).

Additional facts to consider for scRNA-seq

As compared to bulk, scRNA-seq comprises additional steps, including: (i) keeping information of unique molecular identifiers (UMI) to further identify mRNA molecules, and barcodes to identify which mRNA belongs to which single cell, and (ii) reduce experimental biases often triggered by the sequencing of single cells (Table 1). Those biases – mainly known as batch effect and dropout effect – exist for both single-cell and bulk RNA-seq, but are exacerbated by the processing of single cells.

Batch effect encompasses various causes of variation in gene expression estimation, related to technical bias or other independent factors such as variations in sampling, library preparation and sequencing platform, or variations inherent to the origin of the sample (different tumor types). Hence, a frequent challenge is to analyze combined data from various tumor types or from various centers. Batch effect can be corrected *a posteriori* with computational methods, such as Mutual Nearest Neighbors [22] and Canonical Correlation Analysis [23]. Several tools using these and other methods exist, such as ComBat-seq [24],

RUVseq [25] and svaseq [26] for bulk RNA-seq, and kBET [27], Harmony [28], Conos [29] and Seurat [30] for scRNA-seq.

ScRNA-seq, much more than bulk RNA-seq, generates sparse matrices filled with many zeros, corresponding either to a biological feature (truly silent gene) or to a technical artifact, called “dropout effect” and referring to a statistical observation called zero-inflated distribution. The “dropout effect” refers to the limited sampling of RNA content in single-cell wet protocols (falsely silent gene). Indeed, the expression of only 10% of total genes tends to be effectively measured in scRNA-seq experiments [31,32]. Zero-inflated scRNA-seq data can be processed by imputation methods, in order to reduce the impact of the dropout effects, before using pipelines designed for bulk RNA-seq [33,34]. Although deep learning has contributed to significant improvement on this task, the imputation step is rarely performed due to its highly data-dependent efficiency [34,35].

Letting the data speak using statistical analyses of gene expression

Once RNA-seq data are ripe for further analysis, statistical methods are applied to answer different biological questions. Selecting the most appropriate requires bearing in mind the objective of the analysis, the magnitude of the above-cited batch effect and potential statistical biases.

Expression values from bulk RNA-seq can generally be approximated by Poisson, negative-binomial or log-normal distributions and from scRNA-seq by zero-inflated distribution. This limits the use of statistics based on a Gaussian distribution. In practice, simple (log+1)-transformation and expression matrix scaling are often performed, negative-binomial model (adjusted by a regression of bias) [36] can be used, and an increasing number of models now intend to handle zero-inflation in scRNA-seq [37,38]. In the next sections, we review the most common methods used, in a non-exhaustive manner. To illustrate the most common outputs

of RNA-seq analysis (Figure 3), such as volcano plots, we used the published dataset from *Hugo et al*. [39], which comprises bulk RNA-seq data of patients with metastatic melanoma obtained prior to treatment with immune-checkpoint blockers, together with the clinical information of whether or not they responded to treatment. This dataset was primarily used to build an immune signature associated with the response to immune-checkpoint blockers.

Additional facts to consider for scRNA-seq

Expression values from bulk RNA-seq can generally be approximated by Poisson, negative-binomial or log-normal distributions, whereas expression values from scRNA-seq follows a zero-inflated distribution. This limits the use of statistics based on a Gaussian distribution. In practice, simple (log+1)-transformation and expression matrix scaling are often performed, although negative-binomial model (adjusted by a regression of bias) [36] can be used, and an increasing number of models now intend to handle zero-inflation in scRNA-seq [37,38].

Statistics for high-dimensional data: why do we "adjust" the p-value?

RNA-seq data often display a greater number of variables than observations (i.e. more genes than samples analyzed). This high-dimensionality implies a high risk of false discovery. For example, if a statistical test is performed iteratively for each individual gene, and considered significant if p-value <0.05, the user will potentially obtain five false positive results every 100 tested genes. Thus, for 20,000 tested genes, the false positive signals may reach up to 1,000 genes, so the p-values must be adjusted according to the number of tests performed. Multiple methods have been proposed to control this "multiple testing error" as Bonferroni or Benjamini and Hochberg's False Discovery Rate (FDR) [40–42].

Differential expression analysis

Differential expression (DE) analysis aims at identifying significant gene expression variations between two or more experimental groups. For example, DE analysis can be used to uncover genes that are the most differentially expressed in RNA-seq of patients who

responded to immune-checkpoint blockers compared to RNA-seq of patients who did not [25]. DESeq2 [43] and EdgeR [44] are two popular R packages for DE analysis from gene expression data.

Outputs of DE are fold-change of expressions between the different groups, associated with a corrected p-value. The results can be represented using a volcano plot (Figure 3A) or a heatmap (Figure 3B), which highlight the genes that are significantly over- or under-expressed between two or several groups. Volcano plots display the magnitude of both the difference (x-axis) and its statistical meaning (y-axis), while heatmaps use color gradients.

Special aspects of differential expression for scRNA-seq analysis

In scRNA-seq, DE analysis can be used to compare subpopulations of cells or several treatment conditions within a peculiar cell type [45]. Some methods have been developed specifically for scRNA-seq, although their performances are often similar to naive methods or methods originally developed for bulk RNA-seq [46].

Dimensionality reduction

Dimension reduction consists in projecting data from a large space (with many dimensions – for example, each gene being a dimension) into a smaller space (with fewer dimensions – for example, by focusing on genes of interest). Basically, it consists in drastically "summarizing" a dataset to make it more understandable, ideally while retaining most of its intrinsic properties of interest.

Dimension reduction can be carried out using two ways through either feature selection or feature engineering. Eventually, one can apply both approaches sequentially or even iteratively. Feature selection consists in selecting automatically the most important genes for the analysis (for example, the most differentially expressed genes between two conditions, or variable selection with lasso regression, further detailed in Box 1). Feature engineering – or projection – aims at separating samples according to relevant properties in the data. It can

consist in either linear methods – meaning that it preserves all differences between samples – such as Principal Component Analysis (PCA) [47] or Independent Component Analysis (ICA), or non-linear methods such as t-distributed Stochastic Neighbor Embedding (t-SNE) or Uniform Manifold Approximation and Projection (UMAP) (Box 1).

In our example, we used PCA (the most famous linear feature engineering technique) on DE genes (adding feature selection), which reduced the gene list and clearly outlined the two groups of patients, responding or not to immune-checkpoint blockers (Figure 3C). An example of application of non-linear dimensionality reduction is also shown in Figure 3D, using the UMAP method.

Special aspect of dimension reduction for scRNA-seq analysis: trajectory inference

In addition, a rapidly advancing field called "cellular trajectory inference", which uses single cell data, can be considered as a special type of non-linear dimensionality reduction, with the goal to study cellular dynamic processes such as cell cycle, cell differentiation or cell activation. This method – which cannot be applied to bulk data – is based on the assumption that cells that differentiate display a continuous spectrum of states because individual cells will differentiate in an unsynchronized manner (each cell is a snapshot of differentiation time). Tools are therefore used to order the cells along a trajectory based on similarities in their expression patterns, where the "pseudotime" is the unit allowing representing each cell's transcriptional progression toward the terminal state of its trajectory. There are over 75 existing tools for trajectory inference; most of which can be found in the dynverse wrapper [48]. More details regarding trajectory inference are depicted in Box 2.

Clustering

Clustering defines a way to simplify and interpret gene expression data by classifying (or, partitioning) the samples (for bulk RNA-seq) or cells (for scRNA-seq), into groups with gene expression profiles more similar within the group rather than between the groups. Clustering

belongs to the family of unsupervised machine learning methods: groups are inferred by the data themselves, their distribution and their multi-dimensional geometry (and not by predefined classes), among other possible characteristics. Interpretation of the resulting clusters often requires other variables to correlate with, such as phenotypic (type of cancer, toxic exposures, clinical outcome or other continuous data), genetic (for e.g. cluster of gene co-expression with HER2 status in breast cancers), cell types/states (for single-cell data), or technical bias to correct (batch effect). More details are provided in Box 3.

In our RNA-seq example, we have clustered samples based on differentially expressed genes and visualized their relation to response to immunotherapy (Figure 3B, clustering on the left). A second clustering was performed on the same differentially expressed genes in order to identify co/anti-regulated modules (Figure 3B, clustering on the top).

Special aspects of clustering for scRNA-seq analysis

Clustering is a crucial step in single cell data analyses with the objective to characterize the cellular heterogeneity by defining cell types and states. Clustering, (orlike other dimension reduction methods such as PCA, t-SNE or UMAP) can also be used for data analysis and quality control. For example, clustering of a scRNA-seq gene expression matrix may output clusters corresponding to the origin of each cell types identified. In this case, it might be (at least in part) sufficient to characterize the cellular heterogeneity within a sample, for example to appreciate the immune infiltrate composition. Other algorithms such as deconvolution or mixture model-based approach adapted to single-cell data can help adjusting on confounding factors [33,49].

Describing the cancer immune contexture from bulk RNA-sequencing

As we just mentioned, when applied to scRNA-seq data, clustering methods might be sufficient to identify the various immune sub-populations that compose a tumor sample.

Thus, the scRNA-seq technique is probably the most suited for this purpose. However, it is also possible to describe the immune contexture from bulk RNA-seq, and multiple approaches have become available to extrapolate the qualitative and relatively-quantitative compositions of tumor-infiltrating immune cells from bulk data [50,51].

Those methods often rely on prior knowledge of biological processes or cell type specificities. Such information is usually obtained from independent datasets that were established upon purified cell cultures or single-cell experiments.

Enrichment analysis and gene lists applied to immune infiltrate description

Enrichment analysis provides higher-level information compared to gene expression-based analysis. Basically, it enables to identify known biological processes that may be up- or down-regulated within a sample or a population of cells (cluster). Enrichment analysis can therefore also be used to identify and quantify cell type-specific signals, for example to depict whether an identified cluster within a bulk sample is more likely to be CD8+ T cells, macrophages or tumor cells.

Enrichment analysis requires *a-priori* defined sets of genes that are specific to cell types or biological processes. These gene sets are often valid across many biological conditions [52]. For example, ImmuneSigDB is a collection of ~5,000 gene lists derived from ~400 immunological studies, which can be completed with other sources of gene lists [53–56]. Depending on the question in hand, the user may also choose other gene set collections such as MsigDB (Molecular Signatures Database) [57] and KEGG (Kyoto Encyclopedia of Genes and Genomes containing pathways) [58].

Scoring methods are often used to design higher-level variables from the gene expression related to gene sets. Among many scoring methods, Gene Set Enrichment Analysis (GSEA) is the most widely used [57]. GSEA-based approaches rank genes according to their expression variation and then evaluate the enrichment of a particular cell type in a sample,

using a semi-quantitative score. For example, GSEA on bulk RNA-seq data was used to characterize the immune phenotypes of tumor-infiltrating lymphocytes enriched in 598 colorectal cancers [59]. It was also used to build the Cancer Immunome Atlas (https://tcia.at/home) from more than 9,000 samples covering 20 different solid cancers [60]. Enrichment analyses are still under development and recent R packages, such as GAGE [61] and topGO [62], that provide access to several methods in a user-friendly manner.

Deconvolution methods

Deconvolution refers to breaking data up into its various composing elements. Basically, in RNA-seq analysis, the goal of deconvolution is to estimate the relative fractions of individual cell types from a bulk analysis, as a surrogate of what scRNA-seq analysis can produce [50,63]. More than 50 deconvolution algorithms for bulk RNA-seq data are published, with various methods explored, either supervised (the algorithm learns to predict labels from biological measurements) or unsupervised (from the unlabeled data, the algorithm extracts novel features and patterns). Some deconvolution methods assume that each gene expression in a heterogeneous sample is a linear combination of the expression levels of this gene across all the cell types within the sample, weighted by the relative cell fractions [63]. More details are provided in Box 4.

In an alternative approach, scRNA-seq can help deconvolution models applied to bulk RNA-seq: a model can be trained to predict proportions of the different cell types quantified from scRNA-seq, from bulk RNA-seq [64]. This method is notably used by CIBERSORTx [65], an improved version of CIBERSORT, which allows a research of cell signatures, and notably, signatures associated with response to immunotherapy. The joint use of single-cell and bulk RNA-seq for the analysis of a particular tissue or disease can therefore maximize the accuracy and the thoroughness of the final analysis (Table 1).

Clinical considerations, applications and current limitations of RNA-seq

In the era of cancer immunotherapy, RNA-seq technologies can help addressing major biological questions, such as how does the immune system evolve along with cancer progression, how drugs impact anti-tumor immunity and how clinicians can anticipate patients' outcomes and treatment effect, etc (Figure 1). Using a rigorously selected pipeline (from wet to dry lab) among the above-described overall methods, one could theoretically answer a wide variety of biological question from analysis of either bulk or single cell data, albeit with fluctuating performances. With this in mind, we reviewed in the next sections the actual landscape of bulk and scRNA-seq usages in translational and clinical practice.

Predictive biomarkers for immune-checkpoint blockers

Programmed death-ligand 1 (PD-L1) tumoral expression was the first biomarker approved as a companion test for the prescription of the immune-checkpoint inhibitors targeted toward PD-1, such as pembrolizumab. It remains, however, an imperfect biomarker for several reasons, including high rates of false negatives and false positives [66,67], technical issues despite the standardization of immunohistochemistry assays [68,69] and the absence of consensus regarding the relevant staining threshold that should define a PD-L1-positive or –negative tumor [70]. Additionally, tumor mutational load (i.e. the number of mutations per megabase) has been shown to be related to immune-checkpoint blockers efficacy, notably in microsatellite instability-high (MSI-H) tumors [71–78]. In addition to PD-L1 expression, MSI status and tumor mutational burden, an increasing number of immune gene expression signatures are emerging across cancer types, together with methods that aim to estimate neoantigen presentation level, often with the intent to better characterize tumor immune response [39,53,79–81]. Notably, each of these approaches could be addressed using RNA-seq methods. For example, PD-L1 mRNA expression has been shown to correlate both with immunohistochemistry assessment of PD-L1 expression and clinical prediction of response to immune-checkpoint blockade [82,83].

The broad effect of most immunotherapies, including immune-checkpoint blockers, is tempered by the modest proportion of patients who derive a prolonged benefit from it. Up to 30% of patients with commonly sensitive solid tumors (such as melanoma or lung cancer) may in fact be primary refractory to immune-checkpoint blockers [84]. Secondary resistance – a mechanism whereby patient initially responding to a treatment cease to do so – additionally affects up to 20% of patients with melanomas treated in first-line and 30% of patients with lung adenocarcinomas [85]. Finally, about 20% of patients harbor prolonged responses [86]. The type and quantity of immune cells found within the tumor micro-environment influence tumor development and can impact the prognosis of patients [2,87]. The immune infiltrate composition was also shown to be predictive of clinical response to immune-checkpoint blockers [2,88–90]. However, no perfect predictive biomarker has been identified so far and emerging immune signatures often lack validation on independent cohorts, such as in our example. Marker combinations may improve clinical utility. For example, a highly immunogenic tumor – with high PD-L1 expression, CD8+ T cell and dendritic cell infiltration – is more likely to respond to immune-checkpoint blockers than a non-immunogenic tumor [91].

Machine learning for prediction modeling

Machine learning is a branch of artificial intelligence that exploits a large number of statistical techniques to allow mathematical functions to "learn" from experience acquired from similar examples (training). Deep learning is a branch of machine learning stacking several statistical models together to increase nonlinearity between input data and predictions. Machine-learning methods are increasingly used in cancer diagnosis, prognosis and treatment guidance; especially with RNA-seq data as inputs to feed linear models (such as shallow Cox regression) or nonlinear models (such as neural networks). Recent adaptations of neural networks have also integrated methods to learn on survival data (comprising censored values), for example by using a Cox-loss function [92,93].

Prognostic models

Some machine learning approaches have led to clinical applications such as estimating patients' outcome from gene expression values in breast cancers [92,94,95]. Nevertheless, overfitting and poor generalization inherent of high-dimensional data with a small number of examples have limited the success of such methods so far [96,97]. Several teams have developed pan-cancer prognosis predictors by feeding deep learning models with RNA-seq data from the TCGA project [92,93,98,99]. In these studies, concordance indexes (C-indexes) – which provide a global assessment of the model's ability to predict survival – ranged from 0.59 to 0.75, which correspond to modest performance values (0.50 corresponding to random prediction, and 1 to perfect prediction). Limiting overfitting during models training is a very active field of research in machine learning. Several methods exist, that are beyond the spectrum of this review, except for automatic input variable selection such as lasso penalization, which are close to dimension reduction technics detailed above and in Box 1.

Prediction of treatment effect

Evaluation of the effect of treatments has been mostly explored in cell lines [100]. An early study showed that gene expression data was the most efficient data type to predict drug sensitivity, across multi-omics data [100]. Recently, a deep learning-based model based on prior knowledge showed impressive results for the prediction of tumor sensitivity to anticancer compounds [101]. Additionally, *Hwang et al.* recently reported that they were able to accurately predict the response to anti-PD1 treatment in patients with non-small cell lung cancer (with sensitivity, specificity and accuracy of 0.89, 1.0 and 0.95, respectively), by applying a random forest classifier using immune gene signature scores obtained from single-sample GSEA (ssGSEA) [81]. Altogether, an increasing number of gene expression signatures related to response to immunotherapy are emerging from deep and machine learning analyses of RNA-seq data [102,103].

RNA-seq in clinical trials in onco-immunology

Recently, an increasing number of clinical trials have been integrating RNA-seq in their design with various objectives; either biological description of the effect of treatments or with the intent to treat patients (Figure 4). This illustrates the growing interest of clinicians in whole transcriptome profiling. Examples of interventional studies using RNA-seq to guide treatment decisions concerned children with high-grade gliomas (NCT03739372), or adult patients with biliary tract cancers (NCT04318834) or soft-tissue sarcomas (NCT03784014, NCT03375437). Recent evaluations of the use of RNA-seq to orient patients with refractory cancers to targeted treatments demonstrated the feasibility of such an approach in routine; however, it failed to improve patients' outcomes so far [104,105]. Other studies used RNA-seq to evaluate the immune transcriptome profiles of tumors receiving immune-checkpoint blockers, some using immune gene expression signatures (for example in trials NCT03978624, NCT04326257 and NCT03673787).

Studies using scRNA-seq often investigate the changes in tumor immune cell population rates in response to treatments. For example, this has been done in patients receiving endocrine therapy and a CDK4/6 inhibitor for advanced or metastatic breast cancer (NCT04352777), and in patients receiving pembrolizumab with or without chemotherapy for non-small cell lung cancer (NCT04061590).

Yet, the feasibility of the translation of RNA-seq methods into the clinical routine raises many questions, such as the lack of gold standard technical and computational methods. For example, if we consider five types of sequencing libraries, times three different mapping techniques, four imputations, seven normalization processes and four differential expression testings, this leads to 1,680 ways to differentially interpret one dataset. This is largely underestimated if we take into account every parameter that could be manually set at each step. Therefore, standardized pipelines need to be developed to allow reproducibility between centers. Awareness is raising regarding this aspect, and efforts are being made in this direction [106].

Cost

Another important component for evaluating the implementation of RNA-seq into the routine clinical care is cost. The budget for performing RNA-seq depends on many factors, including sequencing depth (average number or reads per transcript), type of library used (rRNA-depleted, poly(A)-enriched RNA, whole RNA sequencing, small RNA sequencing, etc.), type of sequencing (single-end or paired-end), sample size and quality (number of cells to sequence for single-cell), and, if appropriate, cell isolation method (droplet-based, micro-fluidics etc.). Bulk RNA-seq usually requires at least 100-2000 ng of total good quality RNA. This corresponds to 2-10 mg of fresh frozen tissue or 2-4 Formalin-Fixed Paraffin-Embedded (FFPE) slides; although a few companies propose ultra-low input sequencing from 10 pg of high-quality total RNA, at higher costs. Regarding all of these parameters, in 2020, a “typical” experiment of bulk RNA-seq may cost between 150€ and 1500€ per sample; while scRNA-seq still scales between 1000€ and 9000€ per sample (Table 1). These costs do not include bioinformatics services.

Low-cost alternatives are exponentially emerging and intend to reduce the average expense for a whole transcriptome profiling to less than 30€ per sample, mostly by using specific libraries and extensively optimized sequencing techniques [35,107,108]. Additionally, targeted RNA-seq represents an attractive cheaper alternative in case specific transcripts of interest are known beforehand, for example, once a particular immune signature has been built from whole transcriptome analyses and is related to a specific treatment effect. For example, restricting the systematic sequencing to a panel of 500 relevant transcripts could reduce the cost to 10-15€ per sample.

Conclusion

After almost a decade of intensive clinical and translational research in the field on onco-immunology, major questions remain, and notably: can we, should we – and how to – anticipate resistance phenomenon upon immune blockade? Those questions are still pending, yet RNA-seq technologies are on a clear positive path regarding technical feasibility, affordability, the vast number of computational tools associated with an active user community engaged in offering free-access to user-friendly tools. RNA-seq is expected to become a cornerstone of personalized care in onco-immunology. The increasing availability and popularity of innovative techniques further scales up the challenges connected to the analysis of molecular data. Bringing scRNA-seq to the clinic may be challenging, although an increasing number of translational studies should contribute to define its position in cancer treatment personalization.

ScRNA-seq analysis provides expressions of the genes within each cell studied, although recent technological advances allow carrying out several measurements on the same cells. For example, for lymphocytes, the complete variable sequences of the TCR / Ig of each cell can be obtained (Single Cell Immune Profiling). Similarly, the presence of target proteins on the surface of cells can be assessed by CITE-seq [109]. Yet, the major innovation remains the association between gene expression and the location of cell groups in a tissue section, so called the “spatially-resolved transcriptomics” [110,111]. In addition, the single-cell technologies are not restricted to transcriptomics. Single-cell ATAC-seq is well established, and single-cell proteomics are actively under development. Great efforts are being made to integrate “multi-omics” from one cell.

Despite great improvement of the techniques, important issues regarding application of the computational and machine learning-based methods for the analysis of both bulk and scRNA-seq data are emerging. Mainly, it affects reproducibility and interpretability of their results and thus, limits their translationality to the clinic. The most technically advanced approaches, such as artificial neural network-based methods, even if they are very tempting to use, frequently produce results with limited generalization performance. Also, poor human

interpretability keeps on being a limitation, even if no human can interpret clinically the data from a full cancer transcriptome, and while many methods are developed to decipher neural networks decisions. Also, gaining insights in these methods should improve the way we may use it in the clinic. In this context, clinical guidelines for the redaction of clinical trials protocols (SPIRIT-AI) [112] and final reports (CONSORT-AI) [113] of interventions involving artificial intelligence have very recently evolved to take account of these advances.

Finally, single-cell and bulk RNA-seq would both benefit from being developed as complementary techniques (Table 1). In the context of immune therapy, clinical translation of RNA-seq will require consolidating the robustness of predictive features associated with response/resistance to immune-checkpoint blockers. Ultimately, defining standardized analysis pipelines among thousands available will be essential and should benefit to reproducibility and gradual generalization of the practice.

References

1. Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer Immunoediting: Integrating Immunity's Roles in Cancer Suppression and Promotion. *Science* **331**, 1565–1570 (2011).
2. Fridman, W. H., Zitvogel, L., Sautès–Fridman, C. & Kroemer, G. The immune contexture in cancer prognosis and treatment. *Nat. Rev. Clin. Oncol.* **14**, 717–734 (2017).
3. Bruni, D., Angell, H. K. & Galon, J. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. *Nat. Rev. Cancer* **20**, 662–680 (2020).
4. Barnes, T. A. & Amir, E. HYPE or HOPE: the prognostic value of infiltrating immune cells in cancer. *Br. J. Cancer* **117**, 451–460 (2017).
5. Makkouk, A. & Weiner, G. J. Cancer immunotherapy and breaking immune tolerance: new approaches to an old challenge. *Cancer Res.* **75**, 5–10 (2015).
6. Teo, Z. L., Savas, P. & Loi, S. Gene Expression Analysis: Current Methods. in *Molecular Pathology in Cancer Research* (eds. Lakhani, S. R. & Fox, S. B.) 107–136 (Springer, 2016). doi:10.1007/978-1-4939-6643-1_6.
7. Miller, D. F. B. *et al.* Complete Transcriptome RNA-Seq. in *Cancer Gene Networks* 141–162 (Humana Press, New York, NY, 2017). doi:10.1007/978-1-4939-6539-7_10.
8. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* 1–26 (2019) doi:10.1038/s41576-019-0150-2.
9. Heyer, E. E. *et al.* Diagnosis of fusion genes using targeted RNA sequencing. *Nat. Commun.* **10**, 1388 (2019).
10. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
11. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).

12. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
13. Eberwine, J., Sul, J.-Y., Bartfai, T. & Kim, J. The promise of single-cell sequencing. *Nat. Methods* **11**, 25–27 (2014).
14. Fabre, A.-L., Colotte, M., Luis, A., Tuffet, S. & Bonnet, J. An efficient method for long-term room temperature storage of RNA. *Eur. J. Hum. Genet.* **22**, 379–385 (2014).
15. Bossel Ben-Moshe, N. *et al.* mRNA-seq whole transcriptome profiling of fresh frozen versus archived fixed tissues. *BMC Genomics* **19**, 419 (2018).
16. Marczyk, M. *et al.* The impact of RNA extraction method on accurate RNA sequencing from formalin-fixed paraffin-embedded tissues. *BMC Cancer* **19**, 1189 (2019).
17. Turnbull, A. K. *et al.* Unlocking the transcriptomic potential of formalin-fixed paraffin embedded clinical tissues: comparison of gene expression profiling approaches. *BMC Bioinformatics* **21**, 30 (2020).
18. Lafzi, A., Moutinho, C., Picelli, S. & Heyn, H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat. Protoc.* **13**, 2742–2757 (2018).
19. See, P., Lum, J., Chen, J. & Ginhoux, F. A Single-Cell Sequencing Guide for Immunologists. *Front. Immunol.* **9**, 2425 (2018).
20. Andrews, T. S. & Hemberg, M. Identifying cell populations with scRNASeq. *Mol. Aspects Med.* **59**, 114–122 (2018).
21. Fung, C. W., Chan, S. N. & Wu, A. R. Microfluidic single-cell analysis—Toward integration and total on-chip analysis. *Biomicrofluidics* **14**, 021502 (2020).
22. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
23. Canonical Correlation Analysis. in *Applied Multivariate Statistical Analysis* (eds. Härdle, W. & Simar, L.) 321–330 (Springer, 2007). doi:10.1007/978-3-540-72244-1_14.
24. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics Bioinforma.* **2**, (2020).

25. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
26. Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42**, (2014).
27. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).
28. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
29. Barkas, N. *et al.* Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* **16**, 695–698 (2019).
30. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
31. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).
32. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 29 (2016).
33. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1–14 (2018).
34. Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.* **21**, 218 (2020).
35. Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.* **10**, 4667 (2019).
36. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
37. Choi, K., Chen, Y., Skelly, D. A. & Churchill, G. A. Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome Biol.* **21**, (2020).

38. McDavid, A., Gottardo, R., Simon, N. & Drton, M. GRAPHICAL MODELS FOR ZERO-INFLATED SINGLE CELL GENE EXPRESSION. *Ann. Appl. Stat.* **13**, 848–873 (2019).
39. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **165**, 35–44 (2016).
40. Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802 (1988).
41. Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Stat. Med.* **9**, 811–818 (1990).
42. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
43. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
44. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
45. Sandberg, R. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* **11**, 22–24 (2014).
46. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
47. Abdi, H. & Williams, L. J. Principal component analysis. *WIREs Comput. Stat.* **2**, 433–459 (2010).
48. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
49. Huh, R., Yang, Y., Jiang, Y., Shen, Y. & Li, Y. SAME-clustering: Single-cell Aggregated Clustering via Mixture Model Ensemble. *Nucleic Acids Res.* **48**, 86–95 (2020).
50. Finotello, F. & Trajanoski, Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunol. Immunother.* **67**, 1031–1040 (2018).

51. Avila Cobos, F., Vandesompele, J., Mestdagh, P. & De Preter, K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **34**, 1969–1979 (2018).
52. Hoffmann, M. *et al.* Robust computational reconstitution – a new method for the comparative analysis of gene expression in tissues and isolated cell fractions. *BMC Bioinformatics* **7**, 369 (2006).
53. Nirmal, A. J. *et al.* Immune Cell Gene Signatures for Profiling the Microenvironment of Solid Tumors. *Cancer Immunol. Res.* **6**, 1388–1400 (2018).
54. Bard, J., Rhee, S. Y. & Ashburner, M. An ontology for cell types. *Genome Biol.* **6**, R21 (2005).
55. Godec, J. *et al.* Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation. *Immunity* **44**, 194–206 (2016).
56. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218 (2016).
57. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–50 (2005).
58. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
59. Angelova, M. *et al.* Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.* **16**, 64 (2015).
60. Re3data.Org. The Cancer Immunome Atlas. (2016) doi:10.17616/R37D2V.
61. Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. & Woolf, P. J. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10**, 161 (2009).
62. Alexa, A. & Rahnenfuhrer, J. Gene set enrichment analysis with topGO. 26.

63. Shen-Orr, S. S. & Gaujoux, R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr. Opin. Immunol.* **25**, 571–578 (2013).

64. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, (2019).

65. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

66. Ribas, A. *et al.* Association of Pembrolizumab With Tumor Response and Survival Among Patients With Advanced Melanoma. *JAMA* **315**, 1600 (2016).

67. Reck, M. *et al.* Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **375**, 1823–1833 (2016).

68. Ilie, M., Hofman, V., Dietel, M., Soria, J.-C. & Hofman, P. Assessment of the PD-L1 status by immunohistochemistry: challenges and perspectives for therapeutic strategies in lung cancer patients. *Virchows Arch. Int. J. Pathol.* **468**, 511–525 (2016).

69. Ribas, A. & Tumeh, P. C. The future of cancer therapy: selecting patients likely to respond to PD1/L1 blockade. *Clin. Cancer Res.* **20**, 4982–4984 (2014).

70. Aguilar, E. J. *et al.* Outcomes to first-line pembrolizumab in patients with non-small-cell lung cancer and very high PD-L1 expression. *Ann. Oncol.* **30**, 1653–1659 (2019).

71. Rizvi, N. A. *et al.* Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).

72. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207–211 (2015).

73. Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* **371**, 2189–2199 (2014).

74. McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463–1469 (2016).

75. Samstein, R. M. *et al.* Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* **51**, 202–206 (2019).
76. Nebot-Bral, L. *et al.* Hypermutated tumours in the era of immunotherapy: The paradigm of personalised medicine. *Eur. J. Cancer* **84**, 290–303 (2017).
77. Lee, V., Murphy, A., Le, D. T. & Diaz, L. A. Mismatch Repair Deficiency and Response to Immune Checkpoint Blockade. *The Oncologist* **21**, 1200–1211 (2016).
78. Lemery, S., Keegan, P. & Pazdur, R. First FDA Approval Agnostic of Cancer Site — When a Biomarker Defines the Indication. *N. Engl. J. Med.* **377**, 1409–1412 (2017).
79. Galon, J. & Bruni, D. Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. *Nat. Rev. Drug Discov.* **18**, 197–218 (2019).
80. Wang, B. *et al.* Analysis of genetic signatures of tumor microenvironment yields insight into mechanisms of resistance to immunotherapy. *bioRxiv* 2020.03.17.980789 (2020) doi:10.1101/2020.03.17.980789.
81. Hwang, S. *et al.* Immune gene signatures for predicting durable clinical benefit of anti-PD-1 immunotherapy in patients with non-small cell lung cancer. *Sci. Rep.* **10**, 643 (2020).
82. Lau, D., Bobe, A. M. & Khan, A. A. RNA Sequencing of the Tumor Microenvironment in Precision Cancer Immunotherapy. *Trends Cancer* **5**, 149–156 (2019).
83. Conroy, J. M. *et al.* Next generation sequencing of PD-L1 for predicting response to immune checkpoint inhibitors. *J. Immunother. Cancer* **7**, 18 (2019).
84. Hirsch, L., Zitvogel, L., Eggermont, A. & Marabelle, A. PD-Loma: a cancer entity with a shared sensitivity to the PD-1/PD-L1 pathway blockade. *Br. J. Cancer* **120**, 3–5 (2019).
85. Larkin, J. *et al.* Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma. *N. Engl. J. Med.* **373**, 23–34 (2015).
86. Robert, C. *et al.* Three-year overall survival for patients with advanced melanoma treated with pembrolizumab in KEYNOTE-001. *J. Clin. Oncol.* **34**, (2016).
87. Iglesia, M. D. *et al.* Genomic Analysis of Immune Cell Infiltrates Across 11 Tumor Types. *J. Natl. Cancer Inst.* **108**, (2016).

88. Fehrenbacher, L. *et al.* Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. *The Lancet* **387**, 1837–1846 (2016).
89. Teixidó, C., González-Cao, M., Karachaliou, N. & Rosell, R. Predictive factors for immunotherapy in melanoma. *Ann. Transl. Med.* **3**, (2015).
90. Herbst, R. S. *et al.* Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature* **515**, 563–567 (2014).
91. Sharma, P. & Allison, J. P. The future of immune checkpoint therapy. *Science* **348**, 56–61 (2015).
92. Ching, T., Zhu, X. & Garmire, L. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *Plos Comput. Biol.* **14**, e1006076–e1006076 (2018).
93. A, C. & O, G. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinforma. Oxf. Engl.* **35**, i446–i454 (2019).
94. Cardoso, F. *et al.* 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *N. Engl. J. Med.* **375**, 717–729 (2016).
95. Huang, Z. *et al.* Deep learning-based cancer survival prognosis from RNA-seq data: approaches and evaluations. *BMC Med. Genomics* **13**, 41 (2020).
96. Sestak, I. *et al.* Comparison of the Performance of 6 Prognostic Signatures for Estrogen Receptor-Positive Breast Cancer: A Secondary Analysis of a Randomized Clinical Trial. *JAMA Oncol.* **4**, 545–553 (2018).
97. Michiels, S., Koscielny, S. & Hill, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet Lond. Engl.* **365**, 488–492 (2005).
98. Yuan, Y. *et al.* Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol.* **32**, 644–652 (2014).
99. Ramazzotti, D., Lal, A., Wang, B., Batzoglou, S. & Sidow, A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nat. Commun.* **9**, 4453 (2018).

100. Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32**, 1202–1212 (2014).
101. Manica, M. *et al.* Toward Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-Based Convolutional Encoders. *Mol. Pharm.* **16**, 4797–4806 (2019).
102. A, M., S, R., M, W., Dj, S. & D, F. Machine Learning for Cancer Immunotherapies Based on Epitope Recognition by T Cell Receptors. *Front. Genet.* **10**, 1141–1141 (2019).
103. Nb, J. & Av, M. Gene-expression profiling to predict responsiveness to immunotherapy. *Cancer Gene Ther.* **24**, 134–140 (2016).
104. Tuxen, I. V. *et al.* Copenhagen Prospective Personalized Oncology (CoPPO)—Clinical Utility of Using Molecular Profiling to Select Patients to Phase I Trials. *Clin. Cancer Res.* **25**, 1239–1247 (2019).
105. Rodon, J. *et al.* Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. *Nat. Med.* **25**, 751–758 (2019).
106. Tanjo, T., Kawai, Y., Tokunaga, K., Ogasawara, O. & Nagasaki, M. Practical guide for managing large-scale human genome data in research. *J. Hum. Genet.* 1–14 (2020) doi:10.1038/s10038-020-00862-1.
107. Alpern, D. *et al.* BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biol.* **20**, 71 (2019).
108. Mai, Z., Xiao, C., Jin, J. & Zhang, G. Low-cost, Low-bias and Low-input RNA-seq with High Experimental Verifiability based on Semiconductor Sequencing. *Sci. Rep.* **7**, 1053 (2017).
109. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
110. Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat. Methods* **18**, 9–14 (2021).

111. Larsson, L., Frisén, J. & Lundeberg, J. Spatially resolved transcriptomics adds a new dimension to genomics. *Nat. Methods* **18**, 15–18 (2021).

112. Rivera, S. C. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.* **26**, 13 (2020).

113. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374 (2020).

114. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).

115. Sompairac, N. *et al.* Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. *Int. J. Mol. Sci.* **20**, (2019).

116. Nazarov, P. V. *et al.* Deconvolution of transcriptomes and miRNomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients. *BMC Med. Genomics* **12**, 132 (2019).

117. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

118. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4314.

119. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).

120. Chen, H. *et al.* Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat. Commun.* **10**, 1903 (2019).

121. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).

122. Sun, Y., Yao, J., Nowak, N. J. & Goodison, S. Cancer progression modeling using static sample data. *Genome Biol.* **15**, 440 (2014).

123. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

124. Aubin-Frankowski, P.-C. & Vert, J.-P. Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. *Bioinformatics* **36**, 4774–4780 (2020).

125. Specht, A. T. & Li, J. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinforma. Oxf. Engl.* **33**, 764–766 (2017).

126. Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst.* **5**, 251-267.e3 (2017).

127. Xu, Rui. & Wunsch, D. C. *Clustering*. (2008).

128. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).

129. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

130. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).

131. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

132. Regev, A. *et al.* The Human Cell Atlas White Paper. (2018).

133. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–7 (2015).

134. Li, T. *et al.* TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res.* **48**, W509–W514 (2020).

135. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).

136. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).

137. Shen-Orr, S. S. *et al.* Cell type-specific gene expression differences in complex tissues. *Nat. Methods* **7**, 287–9 (2010).

138. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **3**, 346-360.e4 (2016).

139. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6**, e26476 (2017).

140. Monaco, G. *et al.* RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Rep.* **26**, 1627-1640.e7 (2019).

141. Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

142. Petitprez, F. *et al.* The murine Microenvironment Cell Population counter method to estimate abundance of tissue-infiltrating immune and stromal cell populations in murine samples using gene expression. *Genome Med.* **12**, 86 (2020).

143. Chen, Z. *et al.* Inference of immune cell composition on the expression profiles of mouse tissue. *Sci. Rep.* **7**, 40508 (2017).

144. Repsilber, D. *et al.* Biomarker discovery in heterogeneous tissue samples. (2020).

145. H, L., L, S., V, D., O, Y.-H. & W, Z. In silico microdissection of microarray data from heterogeneous cell populations. *BMC Bioinformatics* **6**, 54–54 (2005).

146. Gaujoux, R. & Seoighe, C. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study. *Infect. Genet. Evol.* **12**, 913–921 (2012).

147. Xie, R., Wen, J., Quitadamo, A., Cheng, J. & Shi, X. A deep auto-encoder model for gene expression prediction. *BMC Genomics* **18**, 845 (2017).

148. Dincer, A. B., Janizek, J. D. & Lee, S.-I. Adversarial Deconfounding Autoencoder for Learning Robust Gene Expression Embeddings. *bioRxiv* 2020.04.28.065052 (2020) doi:10.1101/2020.04.28.065052.

Figure legends

Figure 1. Graphical abstract. Clinical positioning of RNA sequencing to personalize onco-immunotherapy treatments from the analysis of a patient's tumor biopsy. The tumor and its microenvironment can be analyzed using gene expression signatures, deconvolution and modeling methods to predict tumor sensitivity to therapies.

Figure 2. Processing pipelines of bulk and single-cell RNA-seq experiments. RNA transcripts are reverse-transcribed into complementary DNA (cDNA), the cDNA second strand is synthesized and then amplified, before being sequenced by a next-generation sequencing (NGS) machine. In the case of scRNA-seq, the reverse transcription step includes the incorporation of unique molecular identifiers (UMI) to identify mRNA molecules, and barcodes to identify single cells. RNA-seq generates 100 nucleotide long reads. A FASTQ file stores all the reads of one sample, in an arbitrary order, together with one quality score per sequenced nucleotide. Mapping (or alignment) refers to finding the position of a read on a reference genome or transcriptome. Quasi-mapping (or pseudo-mapping) aligns reads to a pre-indexed transcriptome rather than to a reference genome and is thus faster. A SAM file stores the information relevant to each read. This information, compressed in binary, is known as a BAM file. Expression estimation relates to counting every read using various normalization methods, and producing gene- or transcript-based quantification fitted for further analysis. Other analyses concerns detection of transcript fusions, splicing variants, or estimate genomic variations, which we do not detail in this review.

Figure 3. Illustration of popular methods used to analyze bulk and single-cell RNA-seq in the field of immune-oncology. For (A) to (C), we used the public GSE78220 dataset of bulk RNA-seq performed on 38 melanoma patients before treatment with immune-checkpoint inhibitors (either pembrolizumab or nivolumab) described in *Hugo et al., Cell (2016)* [25]. (**A**) Volcano plot for differential expression analysis comparing responder and non-responder patients. Gene names are represented in colors for differentially expressed genes with absolute fold change > 0.3 and p-values < 0.01. (**B**) Heatmap organized by hierarchical

clustering of the log transformed gene expression for the differentially expressed genes (defined in A). The row side bar represents responder and non-responder patients. (**C**) PCA of the differentially expressed genes separates responder and non-responder patients in the two first dimensions of the PCA (Dim1 and Dim2). (**D**) UMAP representation of the single-cell RNA-seq analysis of the KUL01-T sample retrieved from the publicly available GSE144735 project (*Ho et al., Nat Genet (2020)*). The sample comprised 1.922 cells analyzed and pre-filtered and 51.292 genes or genomic loci. Expression values have been normalized by an improved negative binomial model. Dimensionality reduction was performed using PCA and then UMAP on the first 30 dimensions, using R Seurat version 3.2.0. The UMAP plot displays the cells (1 point = 1 cell). UMAP efficiently depict the clusters of immune cell types. A gene list of 1149 genes related to activation of invasion and metastasis processes were retrieved (gene list from *Zhang et al., Front Genet (2020)*). Coloring intensity represents the sum of the expressions in the list of genes.

Figure 4. Evolution of the number of opening clinical studies that integrated conventional (A) or single-cell (B) RNA-seq analyses in their designs between 2010 and 2019, according to *clinicaltrials.gov*. Only studies related to cancer were considered. Non-interventional studies are shown in red, interventional studies are shown in black. Dotted lines correspond to estimations based on the number of clinical trials fulfilling the criteria that opened between January 1, 2020 and July 1, 2020.

Table 1. Comparison of bulk and single-cell RNA-seq for clinical translation, specifically in onco-immunology. We also explore the advantages and limitations of combining both techniques.

	RNA-sequencing		
	Conventional bulk	**Single-cell**	**Joint use of single-cell and bulk data**
Experimental and general aspects			
Cost	+	+++	++++
Size of sample (minimal amount of RNA required)	+++	+	+++
Batch effect	+	++	+++
Dropout amplification/coverage bias	+	+++	++
Computational aspects			
Storage capacity	+	++++	++++
Handling sparsity	+	+++	++
Differentiating complex expression patterns from noise	++	+++	++
Dealing with missing data	++	+++	++
Deconvolution requirement to characterize immune infiltrate	++++	+	+
Acquired information			
Appreciating tissue heterogeneity at the cell level	+	++++	++++
Analysing specific populations (T cells, B cells…)	+	+++	++++
Describing populations as a whole	++	++	++++
Accuracy of the final analysis	++	+++	++++

Box 1: To go deeper into dimensionality reduction

Feature selection can consist in the selection of genes from DE analysis by setting thresholds of fold-change and FDR. A reduced gene list comprising only the most differently expressed genes can then be visualized, for example using a heatmap (Figure 3B). Many other approaches of feature selection exist. For example, in the field of supervised machine learning, regularized and/or sparse linear regression (e.g., based on lasso) forces certain coefficients to zero, resulting in genes not used by the model. Non-zero genes are therefore selected as important for the predictions. Unsupervised methods of feature selection can be used by selecting the most fluctuating genes (from estimation of the variance), genes whose dispersion strongly deviates from what would be expected based on their mean expression level (so called over-dispersed genes), [114] or genes strongly contributing to the distinction between data clusters or linear components (see below).

The feature engineering approach to dimensionality reduction consists in transforming the initial data variables, such as gene expression, into a relatively small number of linear or nonlinear combinations. In the case of linearly constructed features, they are frequently referred to as components or meta-genes. The oldest and most widely used linear method for dimensionality reduction is Principal Component Analysis (PCA) [47], which constructs orthogonal sets of vectors in the data space, corresponding to the directions of maximal variance. The classical PCA is a deterministic method while some of its later modifications are probabilistic, e.g., in the case of large single cell datasets. Independent Component Analysis (ICA) is another popular linear method which is usually applied on top of PCA, and creates the most statistically independent features as possible [115]. Both PCA and ICA components can be interpreted either by considering the weights associated to genes (by applying functional enrichment tests) or to samples (confronting with their annotations). In these analyses, ICA components are systematically characterized by improved biological or clinical interpretation [115]. Thus, individual independent components can emerge from the

presence of distinct cell types within a tumor sample: as such, ICA can be used as an unsupervised cell type deconvolution technique [116].

T-distributed Stochastic Neighbor Embedding (t-SNE) [117], and Uniform Manifold Approximation and Projection (UMAP) [118] are non-linear dimensionality reduction methods, based on constructing a pairwise distance similarity graph between data points and representing its structure as faithfully as possible in the reduced 2D or 3D space. t-SNE and UMAP are non-deterministic methods and the resulting axes cannot be directly interpreted. As a matter of fact, t-SNE and UMAP projections can be rotated without their meaning being changed. They can perform significantly better than PCA in grouping on the map points that are close in the large dimensional space, while PCA is more faithful in representing large distances between data points. Compared to t-SNE, the UMAP algorithm better preserves the global structure of the data. Both methods successfully deal with large datasets and are therefore relevant for visualizing scRNA-seq datasets and the results of their analyses. Moderate dimensionality reduction by applying PCA (with 10-50 dimensions being retained) is usually advised before application of t-SNE, which reduces the computational cost and de-noises the single cell profiles to some extent. By contrast, this step is not always mandatory in the case of UMAP, which can be directly applied to large dimensional data.

Box 2: To go deeper into trajectory inference

Trajectory inference is applicable in the case when a population of cells represents a snapshot of an actively developing dynamic process, such as differentiation. In this case, the geometrical structure of the data is recapitulated as a bouquet of diverging and branching trajectories, where branching points represent important cell fate decisions. Each individual cell can be attributed to one or several trajectories and characterized by the value of pseudotime, which reflects the accumulated number of molecular changes as the total path length along the trajectory measured from some root state. The relation between pseudotime and physical time can be highly non-trivial. Dozens of trajectory inference methods such as Monocle [119], STREAM [120], PAGA [121] have been suggested and systematically benchmarked [48]. Interestingly, trajectory-based analysis can be applied to bulk tumoral transcriptomic profiles, too [122]. The usefulness of trajectory inference approaches in onco-immunology is yet to be fully demonstrated.

The trajectory inference is often accompanied by the dynamic network inference whose goal is to discover the gene regulatory networks that drive transitions from one cell type or state to another. Dynamic network inference analysis uses cell ordering from trajectory analysis and co-occurring or correlated genes. Several tools exist (SCENIC [123], GRISLI [124], LEAP [125] and PIDC [126], for example), but the quality of results strongly depends on the dataset.

Box 3: To go deeper into clustering

Clustering based on the transcriptome is attractive given the wide and unbiased nature of the output. However, clustering algorithms contain a crucial parameter that directly or indirectly defines the number of generated clusters – and thus, the resolution or granularity of the analysis. This parameter has a huge impact on downstream conclusions, although it is usually difficult to define before the interpretation step. Thus, underestimating the number of clusters leads to merging important cellular phenotypes together, while overestimating it might make it impossible to identify statistically significant features distinguishing cell states. Therefore, a pragmatic strategy frequently used is to generate several data partitions into clusters at several scales, and develop a multi-scale data analysis strategy.

The most commonly-used algorithms for clustering bulk transcriptomic profiles are hierarchical clustering, k-means and density-based clustering such as DBSCAN [127], and graph-based clustering algorithms also exist such as general-purpose Louvain [128] and Leiden [129]. PhenoGraph [130] approach gained popularity in the single cell data analysis field. These algorithms are based on finding tight communities in point neighborhood graphs which makes them highly scalable and sensitive to relatively fine-grained dissimilarities between cellular transcriptomes [131]. Clustering methods have notably sparked several atlas projects such as the Human Cell Atlas [132].

Using clustering, it is also possible to identify the genes that drive the most the differences between clusters – named “marker genes” – that further reduce the information and help interpretation [50]. In the case of bulk RNA-seq, clustering can output clusters corresponding to tumor types, or to different sequencing techniques. In this case, such analysis can be used as a quality control. If technical (or undesired biological) factors correlate with clusters, additional correction of the batch effect can be considered.

Box 4: To go deeper into deconvolution methods

Recent popular supervised deconvolution algorithms include CIBERSORT [133], MCP-counter [56], TIMER2.0 [134], xCell [135], ESTIMATE [136], csSAM [137], BSEQ-sc [138], EPIC [139] and ABIS [140], among other tools, which focuses on immune-infiltration quantification. A global approach of immune deconvolution could integrate several of these tools to estimate immune cell fractions from bulk RNA-seq data, as proposed by immunedeconv [141]. Although most of these tools only work for human data, some of them -such as mMCP-counter- also offer a version for murine data [142] (adapted from MCP-counter) and ImmuCC [143] (derived from CIBERSORT).

Unsupervised deconvolution algorithms, often based on matrix factorization methods [116,144–146], work without predefined reference cell type signatures. Recent unsupervised methods have used neural networks in the form of auto-encoders, where the thinnest hidden layer of the neural network is extracted to represent a lower-dimension of the data [147,148].

Declaration of Interest statement

MK, DM, MA, FXD and AZ have no conflict of interest to declare. LV reports personal fees from Adaptherapy, non-personal fees from Pierre-Fabre and Servier, grants from Bristol-Myers Squibb, all outside the submitted work. As part of the Drug Development Department (DITEP), AM and LV report being: Principal/sub-Investigator of Clinical Trials for Abbvie, Adaptimmune, Aduro Biotech, Agios Pharmaceuticals, Amgen, Argen-X Bvba, Arno Therapeutics, Astex Pharmaceuticals, Astra Zeneca Ab, Aveo, Basilea Pharmaceutica International Ltd, Bayer Healthcare Ag, Bbb Technologies Bv, Beigene, Blueprint Medicines, Boehringer Ingelheim, Boston Pharmaceuticals, Bristol Myers Squibb, Ca, Celgene Corporation, Chugai Pharmaceutical Co, Clovis Oncology, Cullinan-Apollo, Daiichi Sankyo, Debiopharm, Eisai, Eisai Limited, Eli Lilly, Exelixis, Faron Pharmaceuticals Ltd, Forma Tharapeutics, Gamamabs, Genentech, Glaxosmithkline, H3 Biomedicine, Hoffmann La Roche Ag, Imcheck Therapeutics, Innate Pharma, Institut De Recherche Pierre Fabre, Iris Servier, Janssen Cilag, Janssen Research Foundation, Kura Oncology, Kyowa Kirin Pharm. Dev, Lilly France, Loxo Oncology, Lytix Biopharma As, Medimmune, Menarini Ricerche, Merck Sharp & Dohme Chibret, Merrimack Pharmaceuticals, Merus, Millennium Pharmaceuticals, Molecular Partners Ag, Nanobiotix, Nektar Therapeutics, Novartis Pharma, Octimet Oncology Nv, Oncoethix, Oncopeptides, Orion Pharma, Ose Pharma, Pfizer, Pharma Mar, Pierre Fabre, Medicament, Roche, Sanofi Aventis, Seattle Genetics, Sotio A.S, Syros Pharmaceuticals, Taiho Pharma, Tesaro, Xencor. Research Grants from Astrazeneca, BMS, Boehringer Ingelheim, Janssen Cilag, Merck, Novartis, Onxeo, Pfizer, Roche, Sanofi. Non-financial support (drug supplied) from Astrazeneca, Bayer, BMS, Boringher Ingelheim, Medimmune, Merck, NH TherAGuiX, Onxeo, Pfizer, Roche. AZ was supported by the Ministry of Science and Higher Education of the Russian Federation (Project No. 075-15-2020-808) and by European Union's Horizon 2020 program (grant No. 826121, iPC project).

Acknowledgment

We would like to thank colleagues involved in RNA-seq analysis and molecular analysis of tumor immunity for inspiring talks on this topic, including: Guillaume Meurice, Urszula Czerwinska, Nicolas Sompairac, Laurence Zitvogel, Christophe Massard, Antoine Hollebecque.

Figure 1 (graphical abstract)

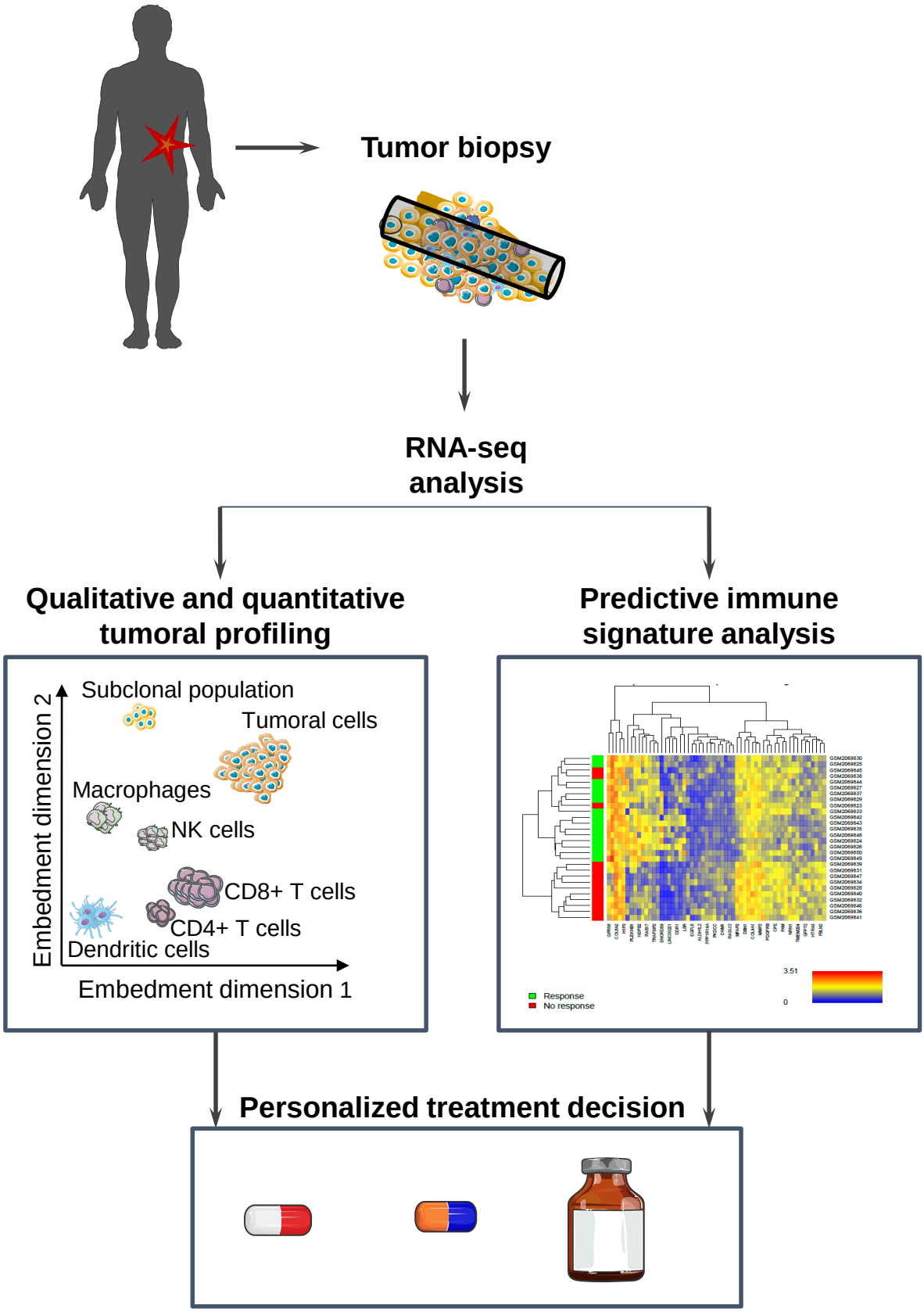


Figure 2

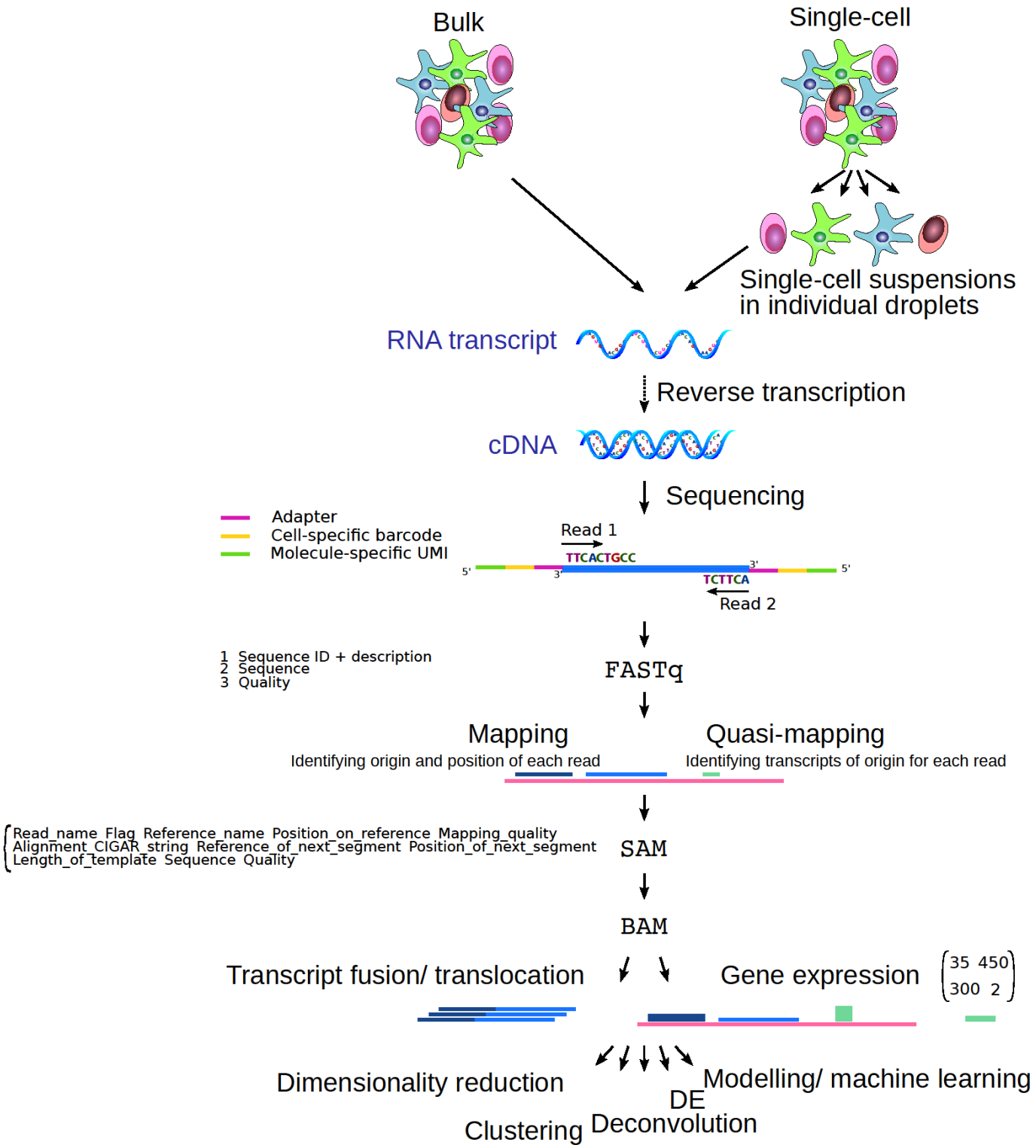
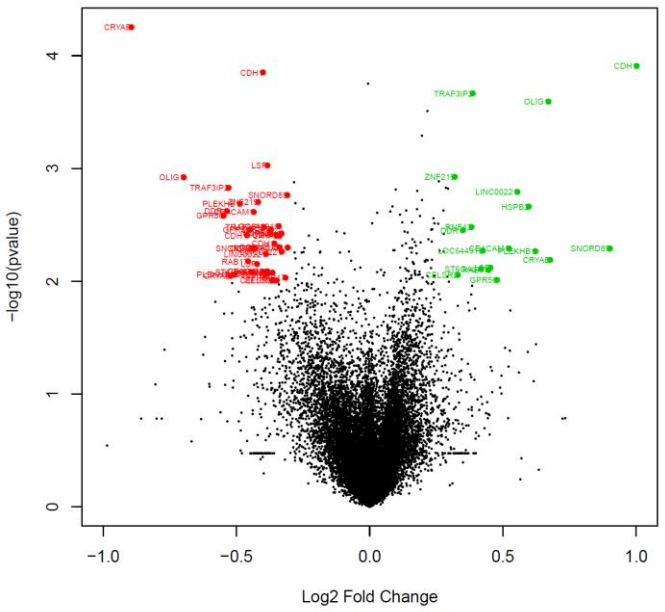

Bulk
Single-cell
Single-cell suspensions in individual droplets
RNA transcript
Reverse transcription
cDNA
Sequencing
Adapter
Cell-specific barcode
Molecule-specific UMI
Read 1
TTCACTGCC
TCTTCA
Read 2
1 Sequence ID + description
2 Sequence
3 Quality
FASTq
Mapping
Identifying origin and position of each read
Quasi-mapping
Identifying transcripts of origin for each read
Read_name Flag Reference_name Position_on_reference Mapping_quality
Alignment_CIGAR_string Reference_of_next_segment Position_of_next_segment
Length_of_template Sequence Quality
SAM
BAM
Transcript fusion/ translocation
Gene expression
35 450
300 2
Dimensionality reduction
Clustering
DE
Deconvolution
Modelling/ machine learning

Figure 3

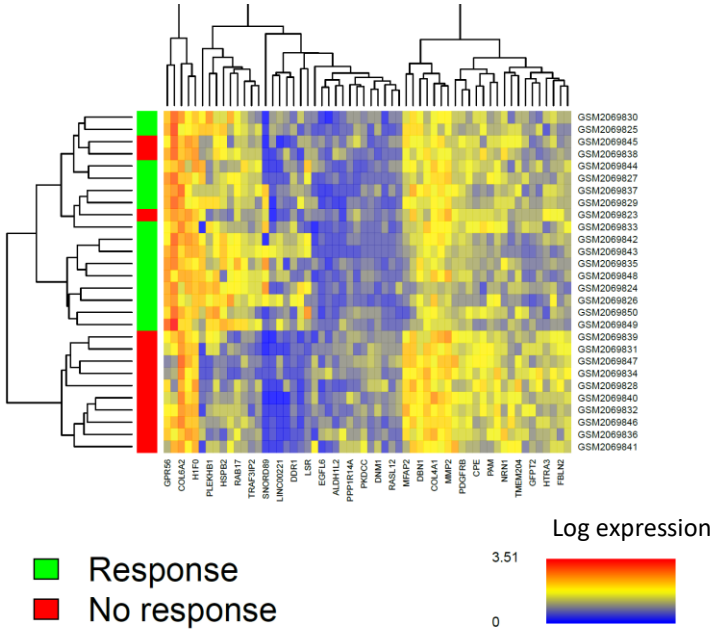
A

Volcano plot for differential expression



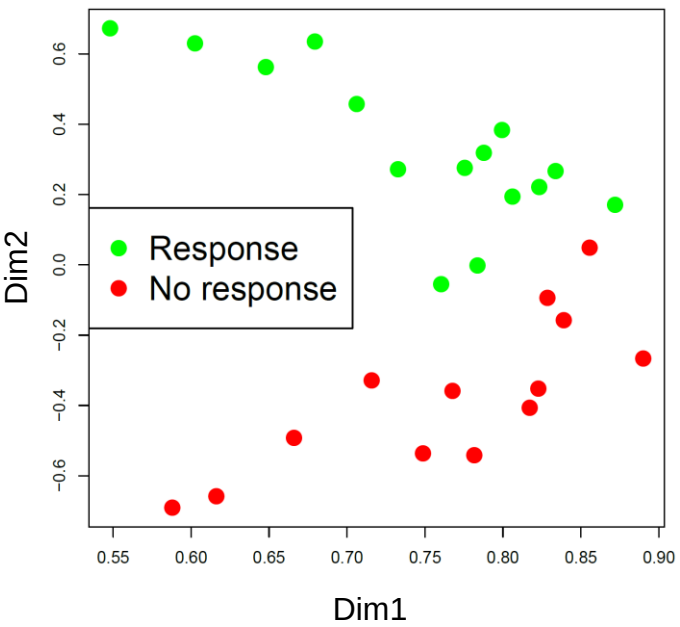
B

Heatmap of the DE genes



C

PCA for DE genes



D

UMAP for sub-population segregation

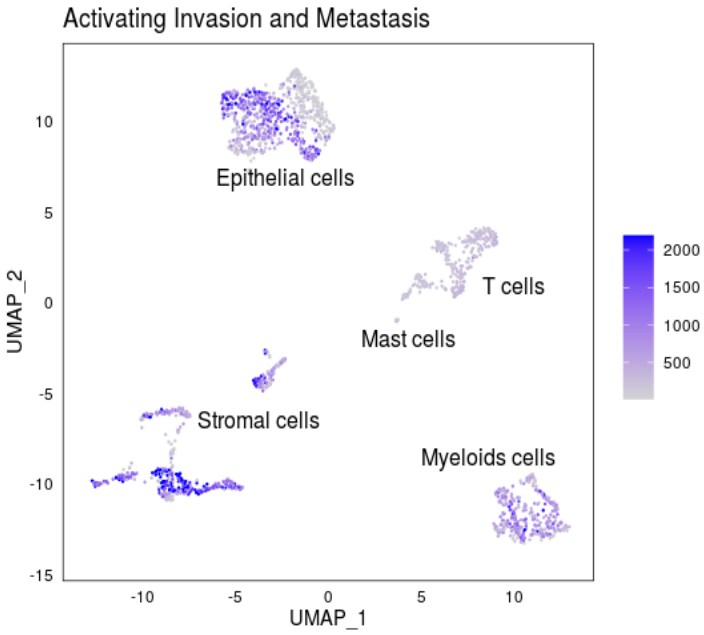
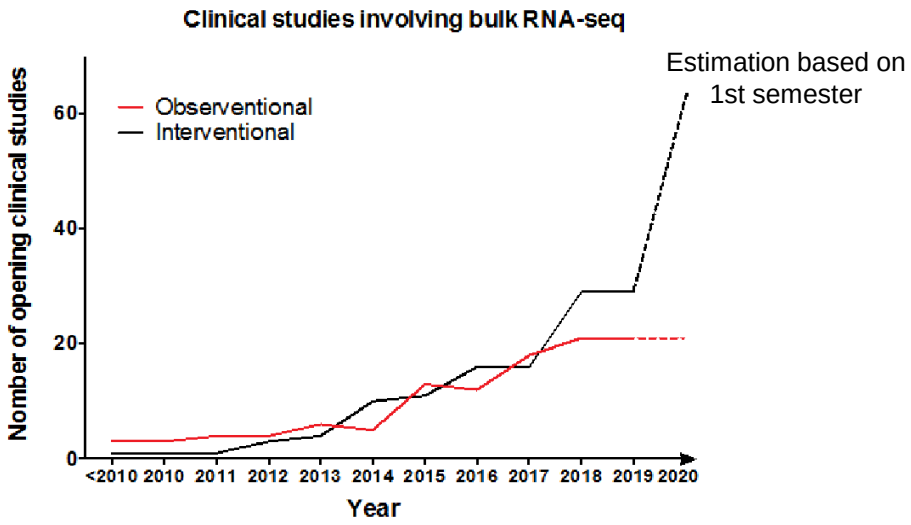


Figure 4

A


Clinical studies involving bulk RNA-seq
Nomber of opening clinical studies
Observentional
Interventional
Estimation based on
1st semester
60
40
20
0
<2010 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
Year

B

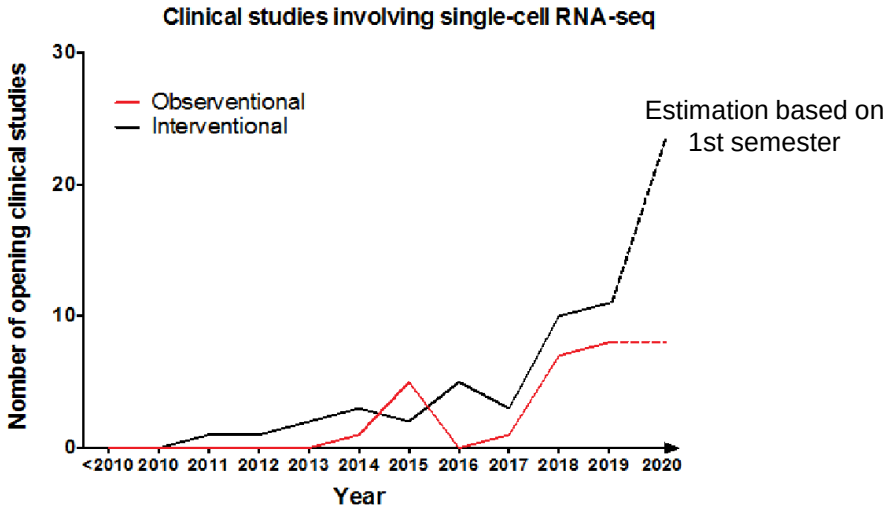

Clinical studies involving single-cell RNA-seq
Nomber of opening clinical studies
Observentional
Interventional
Estimation based on
1st semester
30
20
10
0
<2010 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
Year

Figure 1 (graphical abstract)